文章编号: 1005-3026(2003)04-0327-03

中文 WordNet 的研究及实现

张 俐, 李晶皎, 胡明涵, 姚天顺
(东北大学 信息科学与工程学院, 辽宁 沈阳 110004)

摘 要: 提出了一种从英文 WordNet 转换生成中文 WordNet 的方法,并设计实现了中文 WordNet 的转换生成系统.论述了在构造中文 WordNet 的语义网络时,概念结点的转换原则,中文词形与词义映射关系的重新聚合,以及转换生成中文 WordNet 的可行性及转换中的相关问题,并给出了一种依据 WordNet 进行节点转换自消岐的方法.

关 键 词: WordNet;中文 WordNet;同义词集合;概念映射
中图分类号: TP 391 **文献标识码**: A

为使计算机具有理解和处理自然语言的能力,必须使计算机拥有词法、句法、语义、语用、语境等丰富的语言知识.随着自然语言处理和机器翻译技术的发展,对于多层次、多类型、多关系的语义体系的需求越加紧迫.

1 WordNet 与中文 WordNet

英文 WordNet 是普林斯顿大学认知科学实验室开发的一部在线词典数据库系统,是基于英文的词汇语义网络系统.WordNet 将英文的名词、动词、形容词和副词组织为同义词集合(synsets),每一个集合表示一个基本的词汇概念,并在这些词汇概念间建立了包括同义关系、反义关系、上位关系、下位关系、部分关系以及完全关系等多种词汇语义关系[1~3].目前,WordNet 被成功地用于词义消歧、语言学自动处理、双语及多国语机器翻译、检索系统等一系列语言工程,被普遍认为是用于计算语言学、文本分析和许多相关领域的最重要的资源[4],在国际计算语言学界已有相当的影响.2001 年,成立了 WordNet 研究学会,2002 年 2 月于印度召开了第一届 WordNet 国际会议.许多国家都已着手实施构造本民族语言的 WordNet,荷兰、西班牙、意大利、英国、法国、德国、捷克、爱沙尼亚等国家都参与了构建 EuroNet 系统.此外,以 WordNet 为基础,韩国也已经开始构造韩文 WordNet(KoreaNet)系统,初步的成果已经在韩英机器翻译系统中得到了验证[5~7].

汉语的语法特征较弱,语义知识,特别是词汇语义知识对于汉语的信息处理就显得更为重要[8].本文对 WordNet 进行了较为深入的分析研究,通过转换生成的方法初步实现了中文 WordNet 的半自动建造,期望能够成为与国际计算语言界接轨的切入点.

2 中文 WordNet 的建造

2.1 中文 WordNet 的实现方法

以 WordNet 中的概念间关系为基础,以半自动方式创建一个适用于中文信息处理的中文 WordNet 系统.首先抽取 WordNet 的语义框架及概念间关系,将其移植为中文 WordNet 框架,然后在这个框架中植入汉语词汇,自顶向下的转换生成中文 WordNet.这样,就可以保证在较高的概念层次上与 WordNet 兼容,在较低的概念层次上具有最大限度的灵活性.

2.2 实现中的几个问题

(1) 从结构上讲,词汇语义体系具有层次性、网络性和开放性.WordNet 是计算机科学家与心理语言学家十几年心血的结晶,尽管不同民族语言之间存在着差异,但从概念角度来说,人们对世界的认识还是相通的、相近的、甚至是相同的[9].下面是 WordNet 中事件类名词概念以及其直接下位概念.

收稿日期: 2002-10-22
基金项目: 国家自然科学基金资助项目(69985001).
作者简介: 张 俐(1961 -),女,辽宁沈阳人,东北大学副教授; 李晶皎(1963 -),女,辽宁沈阳人,东北大学教授; 姚天顺(1934 -),男,上海人,东北大学教授,博士生导师.

5.3 事件

5.3.1 相伴事件	5.3.2 雪崩
5.3.3 经验,阅历	5.3.4 灾难,事故,困难
5.3.5 奇迹,奇事	5.3.6 奇观
5.3.7 事,事情	5.3.8 插曲
5.3.9 不测,意外	5.3.10 开始,开端
5.3.11 终止,结束	5.3.12 周期性事件
5.3.13 变化,转变	5.3.14 偶然事件
5.3.15 起火,着火	5.3.16 事变
5.3.17 释放能量	5.3.18 实例,例子,范例
5.3.19 运动,移动	5.3.20 失败
5.3.21 成功,胜利	5.3.22 出现,露面
5.3.23 命运,天数	5.3.24 消失
5.3.25 接触,碰撞	5.3.26 中断,破裂
5.3.27 发音,出声	5.3.28 联合,结合
5.3.29 新闻事件	5.3.30 闪现
5.3.31 集中	5.3.32 时机,关头
5.3.33 突发,爆发	5.3.34 发作
5.3.35 逆转,倒退	5.3.36 繁荣
5.3.37 破产,倒闭	

与同义词词林[10]中事件类名词概念对比,两者有十分相似的基本概念.相比之下,WordNet 的继承结构层次更深.因此,WordNet 的概念结构框架以及大多数词汇语义关系都可以保留在中文 WordNet 中.

(2) WordNet 本质上是一个词汇概念网络,描述的是概念间的各种语义关系[11,12].而事实上,远非所有的概念在语言中都有相对应的词汇,特别是在词汇等级的最顶层或较高的层次上.因此,在各种语言中都存在着这样的概念,它们是用短语或复合词来描述的[13].WordNet 中就定义了大量的复合词,用来表示英语中未曾词汇化的那些概念结点.在转换中,常常遇到这样的概念结点映射,如 tour de force| 绝技,free-living| 无拘无束, secret ballot| 无记名投票;而 intercommunion| 各教派间举行的圣餐,是将中文中没有的概念用汉语的短语来描述的例子.

(3) 不同语言之间毕竟存在着差异.例如,英语中的一个范畴在汉语中会划分为不同的范畴:穿/ to put on (of garments),戴/ to put on (of other things such as caps, armbands, etc.);而英语中分为两种范畴的词语在汉语中可以视为一种范畴:table(桌子),desk(书桌),cup(杯子),mug(有柄的大杯),glass(玻璃杯),等等.还有一些汉语中的概念是 WordNet 中不存在的.因此,转换生成的中文 Wordnet 还需进一步作必要的调整、合并或扩充,并重新聚合汉语词形与概念结点的映射关系.所有上述这样的结点往往是处于较低层次上的叶子结点,对于它们的调整不会影响到 WordNet 的框架结构,因此,从整体来说,并不影响两个语种的 WordNet 之间的映射和融合.

3 中文 WordNet 的生成系统

3.1 语义网络的转换

多义词和同义词是自然语言中普遍存在的现象,不同的语言,其词形与词义之间的映射关系是不同的.涉及到同义词集合的转换,必须处理一对多、多对一及多对多等多种转换关系.因此,转换生成中文 WordNet 的核心问题,是如何将词网中以英文同义词集合表示的概念结点,准确地转换为中文同义词集合表示,以及怎样重新聚合中文词形与词义的映射关系.

3.2 生成系统的结构

WordNet 转换系统框图如图 1.其中,虚线内的两个功能模块可以根据不同时期的需要随时分别加载或卸载.

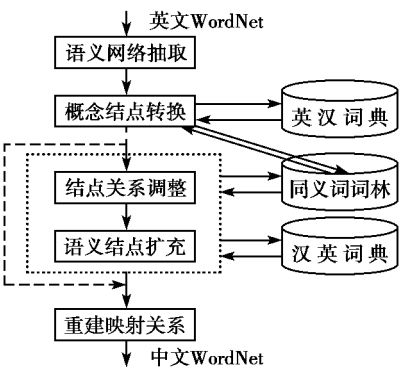

图 1 WordNet 转换系统

Fig. 1 The transformation system of WordNet

4 概念结点转换中的消歧算法

在概念结点转换的过程中,不可避免地要遇到词汇消歧的问题,而 WordNet 本身就是一种可以用来解决歧义的资源.

4.1 两种 WordNet 的概念节点间的基本关系

设函数 $\delta(x)$ 是从英文到中文的单词映射关系,则 $\delta(x)=\{y_i \mid 1\leq i\leq n\}$;其中 x 代表英文单词,y 则表示中文单词.同义词集合 $s\_En=\{En_1, En_2, \cdots, En_m\}$ 表示 WordNet 中的一个概念节点,同义词集合 $s\_Ch=\{Ch_1, Ch_2, \cdots, Ch_m\}$ 为 s_En 映射到中文 WordNet 中的概念节点,则有:

$$s\_Ch=\delta(En_1)\cap\delta(En_2)\cdots\cap\delta(En_n)=\{Ch \mid Ch\in\delta(En_i)\cap En_i\in s\_En; 1\leq i\leq n\}.$$

上面 s_Ch 表示的是各英文之中文对译词的词形的交集,实际转换中,应考虑取各对译词之词义的交集作为 s_Ch.

4.2 消岐算法

step 1　对于同义词集合 $s_- En$ 中的一个英文单词 En,查找英汉辞典,生成中文对译词集合 $Ch = \{ ch_1, ch_2, \cdots \}$.

step 2　访问同义词词林,对上述 Ch_i 中的各中文单词,进行语义分类的相似计算,并依据相似度划分出义项子集,为 $Ch = \{ C_1, C_2, \cdots, C_i, \cdots \}$;其中,$C_i = \{ ch_{i1}, ch_{i2}, \cdots \}$ 是一个中文同义词集合,对应着英文单词 En 的第 i 个义项.

step 3　针对词形集合 $Ch = \{ C_1, C_2, \cdots, C_i, \cdots \}$,建立对应的语义分类码集合 $S = \{ s_1, s_2, \cdots, s_i, \cdots \}$;其中,$s_i$ 为义项 C_i 对应的语义分类码.

step 4　重复上述步骤,直至求出 $s_- En = \{ En_1, En_2, \cdots, En_m \}$ 中每一个英文单词 En_i 对应的语义分类码集合 $S_i = \{ s_{i1}, s_{i2}, \cdots, s_{ij}, \cdots \}$.

step 5　计算 $s = S_1 \cap S_2 \cdots \cap S_n$;对于每一个英文单词 En_i 对应的 $Ch_i = \{ Ci_1, Ci_2, \cdots \}$;生成集合 $syn_- Ch_i = \{ C \mid C \in Ch_i, C$ 所对应的语义分类码为 $s \}$.

step 6　$s_- Ch = syn_- Ch_1 \cup syn_- Ch_2 \cup \cdots \cup syn_- Ch_n$.

5 结　论

(1) 抽取得到了 WordNet 的名词和动词的全部词汇语义框架和概念关系,并针对上述名词和动词近 90 000 个概念结点进行了节点自动转换.名词转换的成功率达 80 %,正确率为 50 %.还设计实现了相应的工具软件,对 8 000 余个名词概念结点以及约 5 000 个动词结点进行了后校正.

(2) 语义体系转换工作是动态进行的,转换后的结果可以直接在 WordNet 环境下运行,图 2 是中文 WordNet 的运行实例.

(3) 将对 WordNet 中的形容词和副词进行转换,并针对形容词和副词的特点,改进算法,提高这一部分自动转换的准确性.

图 2　"减员"的上位关系

Fig. 2　Hypernyms of a noun "减员"

参考文献:

[1] Miller G A. An on-line lexical database[J]. *International Journal of Lexicography*, 1990, 3(4): 235 - 244.

[2] Gross D, Miller K. Adjectives in WordNet[J]. *International Journal of Lexicography*, 1990, 3(4): 265 - 277.

[3] Fellbaum C. Co-occurrence and antonymy[J]. *International Journal of Lexicography*, 1995, 8(4): 281 - 303.

[4] Fellbaum C, Miller G A, Curtiss S, *et al*. An auditory processing deficit as a possible source of SLI [A]. *Proceedings of the 19th Boston University Conference on Language Development* [C]. Ithaca, NY: Cascadilla Press, 1995. 204 - 215.

[5] Fellbaum C. A lexical database of english: the mother of all WordNets[A]. *EuroWordNet*[C]. Holland: Kluwer, 1998. 137 - 148.

[6] Moon Y. Design and implementation of WordNet for Korean nouns [J]. *Journal of the Korea Information Science Society*, 1996, 2(4): 437 - 445.

[7] Dorr B J. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation [J]. *Machine Translation*, 1997, 12(1): 1 - 55.

[8] 邱广君,张俐. 汉语信息处理中的语义关系类型分析[J]. 东北大学学报(自然科学版), 1998, 19(1): 48 - 51.
(Qiu G J, Zhang L. A analysis of the semantic relation in Chinese information processing[J]. *Journal of Northeastern University (Natural Science)*, 1998, 19(1): 48 - 51.)

[9] 杰弗里·N·利奇. 语义学[M]. 上海: 上海外语教育出版社, 1987. 325 - 357.
(Leech G N. *Semantics*[M]. Shanghai: Foreign Language Education Press of Shanghai, 1987. 325 - 357.)

[10] 梅家驹. 同义词词林[M]. 上海: 上海辞书出版社, 1996. 108 - 166.
(Mei J J. *A synonyms dictionary*[M]. Shanghai: Shanghai Thesaurus Press, 1996. 108 - 166.)

[11] Fellbaum C. The organization of verbs and verb concepts in a semantic net[A]. *Predicative Forms in Natural Language and in Lexical Knowledge Bases*[C]. Holland: Kluwer, 1998. 93 - 109.

[12] Beckwith R, Miller G A. Implementing a lexical network [J]. *International Journal of Lexicography*, 1990, 3(4): 302 - 312.

[13] 石安石. 语义论[M]. 北京: 商务印书馆, 1993. 87.
(Shi A S. *Semantics*[M]. Beijing: The Commercial Press, 1993. 87.)

Implementation of Chinese WordNet

ZHANG Li, LI Jing-jiao, HU Ming-han, YAO Tian-shun
(School of Information Science & Engineering, Northeastern University, Shenyang 110004, China. Correspondent: ZHANG Li, associate professor, E-mail: Zhang_Li0501@sina.com)

Abstract: An approach based on transformation from WordNet to Chinese WordNet is proposed. The system of the Chinese WordNet has been implemented. Some rules for synset's translating and some methods for reconstituting relation between Chinese word form and word meaning are given. The feasibility and relative problems of the Chinese WordNet are discussed. A self-disambiguation algorithm for the transformation of concept nodes is prescribed.

Key words: WordNet; Chinese WordNet; synsets; concept map

(*Received October* 22, 2002)